Floor Discussion

John Priebe
Bureau of the Census

The first paper, Automatic coding of Handprinted Responses in Statistical Surveys, was presented by Svein Nordbotten of the University of Bergen, Norway. Hand printed responses to censuses and surveys need to be coded by the statistical agency before tabulations can take place. Modern optical scanning equipment can transcribe the answers to machine-readable form. Because of distorted characters and unusual spelling, automatic coding may not always be an easy solution. The paper outlines an automated model for handling distorted characters and spelling errors. The model was tested on a list of U.S. cities over 50,000 population by means of a prototype system. The system used an 8 by 8 matrix for each of the 26 letters in the alphabet. A number of tests were run with samples of increased distortion of letter shapes or missing letters. Most of the tests had good results, but the most distorted letters were often not recognized.

The second paper, SICORE, The INSEE Automatic Coding System, was presented by Pierrette Schuhl of INSEE, the French National Institute of Statistics. She was assisted by Pascal Riviere who launched the project in 1993. The SICORE project aimed to improve and generalize QUID which had been used for more than 10 years. It will soon be completed, and is in use or being tested on some surveys. The four main objectives were to construct the knowledge base for variables to be coded, create an adequate management structure, write a generalized software package, and provide a documented methodology.

The first paper:

Q: Does your algorithm recognize false positives? Did you have many cases where a city was mis-identified?

A. My system always comes to a conclusion. Correct city names were not always assigned to a response. Generally the quality is better than manual coding. This was determined by having experts recode the files and comparing the results.

The second paper:

Q. by Dr. Wouter Keller (Netherlands): When one looks at search engines on the Internet, most use tri-grams with overlapping fragments. Did you use overlapping fragments?

A: No, I did not. The reason is that the tree structure makes it impossible to use overlapping fragments.

Q: When you use the tree structure, if the first letter of a response is missing, you have very heavy loss of recognition, as apposed to missing letters at the elsewhere in the response. For example if "d" is missing in "doctor." Did you do anything to compensate for this.

A: That is correct, but we found that this did not happen often. If one finds many occurrences of the first letter missing for a particular word, that word can be put into the knowledge file. In this example "octer" would be in the file, if it occurred often.

Q: Table 2 shows that for four digit occupation codes, 80 percent of the responses were coded with an 8 percent error rate. If the 8 percent was concentrated in one area, it may be very serious. How have you addressed this problem?

A: This is why the SICORE loop is so important. After a file is coded a sample is recoded by experts. If a particular text was often coded poorly, we can find it.

Q: Can you lower coding efficiency and get better accuracy?

A: We can chose between better efficiency or better accuracy. After coding, SICORE checks a set of parameters. User can chose the parameters that will lower efficiency, but increase accuracy.